Interrater Reliability of the Structured Clinical Interview for the DSM-5 Alternative Model of Personality Disorders Module I

Buer Christensen, Tore; Paap, Muirne C. S.; Arnesen, Marianne; Koritzinsky, Karoline; Nysaeter, Tor-Erik; Eikenaes, Ingeborg; Germans Selvik, Sara; Walther, Kristoffer; Torgersen, Svenn; Bender, Donna S.

Published in:
Journal of Personality Assessment

DOI:
10.1080/00223891.2018.1483377

Publication date:
2018

Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: http://www.tandfonline.com/loi/hjpa20

Interrater Reliability of the Structured Clinical Interview for the *DSM–5* Alternative Model of Personality Disorders Module i: Level of Personality Functioning Scale

Tore Buer Christensen, Muirne C. S. Paap, Marianne Arnesen, Karoline Koritzinsky, Tor-Erik Nysaeter, Ingeborg Eikenaes, Sara Germans Selvik, Kristoffer Walther, Svenn Torgersen, Donna S. Bender, Andrew E. Skodol, Elfrida Kvarstein, Geir Pedersen & Benjamin Hummelen

To cite this article: Tore Buer Christensen, Muirne C. S. Paap, Marianne Arnesen, Karoline Koritzinsky, Tor-Erik Nysaeter, Ingeborg Eikenaes, Sara Germans Selvik, Kristoffer Walther, Svenn Torgersen, Donna S. Bender, Andrew E. Skodol, Elfrida Kvarstein, Geir Pedersen & Benjamin Hummelen (2018) Interrater Reliability of the Structured Clinical Interview for the *DSM–5* Alternative Model of Personality Disorders Module i: Level of Personality Functioning Scale, Journal of Personality Assessment, 100:6, 630-641, DOI: 10.1080/00223891.2018.1483377

To link to this article: https://doi.org/10.1080/00223891.2018.1483377

JOURNAL OF PERSONALITY ASSESSMENT
2018, VOL. 100, NO. 6, 630–641
https://doi.org/10.1080/00223891.2018.1483377

Routledge
Taylor & Francis Group

INSTRUMENT DEVELOPMENT

Interrater Reliability of the Structured Clinical Interview for the *DSM–5* Alternative Model of Personality Disorders Module i: Level of Personality Functioning Scale

Tore Buer Christensen[1], Muirne C. S. Paap[2], Marianne Arnesen[3], Karoline Koritzinsky[3,*], Tor-Erik Nysaeter[1], Ingeborg Eikenaes[4], Sara Germans Selvik[5,6], Kristoffer Walther[7], Svenn Torgersen[3], Donna S. Bender[8], Andrew E. Skodol[9], Elfrida Kvarstein[7], Geir Pedersen[7,10] and Benjamin Hummelen[7,11]

[1]Department of Mental Health, Sorlandet Hospital, Arendal, Norway; [2]Department of Special Needs, Education, and Youth Care, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, The Netherlands; [3]Department of Psychology, University of Oslo, Oslo, Norway; [4]Department of Personality Psychiatry, Division of Mental Health and Addiction Treatment, Vestfold Hospital Trust, Vestfold, Norway; [5]Department of Psychiatry, Hospital Namsos, Namsos, Norway; [6]Department of Mental Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway; [7]Department of Personality Psychiatry, Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway; [8]Department of Psychiatry and Behavioral Sciences and Counseling and Psychological Services, Tulane University; [9]Section of Personality Psychiatry, Clinic Mental Health and Addiction, Oslo University Hospital, Oslo, Norway; [10]NORMENT, KG Jebsen Center for Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway; [11]Department of Research and Development, Clinic Mental Health and Addiction, Oslo University Hospital, Oslo, Norway

ABSTRACT

The fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–5*) presents an alternative model for personality disorders in which severity of personality pathology is evaluated by the Level of Personality Functioning Scale (LPFS). The Structured Interview for the DSM–5 Alternative Model for Personality Disorders, Module I (SCID–5–AMPD I) is a new tool for LPFS assessment, but its interrater reliability (IRR) has not yet been tested. Here we examined the reliability of the Norwegian translation of the SCID–5–AMPD I, applying two different designs: IRR assessment based on ratings of 17 video-recorded SCID–5–AMPD I interviews by five raters; and test–retest IRR based on interviews of 33 patients administered by two different raters within a short interval. For the video-based investigation, intraclass correlation coefficient (ICC) values ranged from .77 to .94 for subdomains, .89 to .95 for domains, and .96 for total LPFS. For the test–retest investigation, ICC ranged from .24 to .72 for subdomains, .59 to .90 for domains, and .75 for total LPFS. The test–retest study revealed questionable reliability estimates for some subdomains. However, overall the level of personality functioning was measured with a sufficient degree of IRR when assessed by the SCID–5–AMPD I.

ARTICLE HISTORY
Received 11 September 2017
Revised 7 May 2018

The *Diagnostic and Statistical Manual of Mental Disorders* (*DSM;* American Psychiatric Association, 1980) first recommended personality disorder (PD) classification and definition based on explicit diagnostic criteria, representing an important step toward the systematic assessment of personality psychopathology. However, the current categorical model has severe shortcomings, including extensive comorbidity among PD categories (Grant, Stinson, Dawson, Chou, & Ruan, 2005; Torgersen, Kringlen, & Cramer, 2001; Zimmerman, Chelminski, Young, Dalrymple, & Martinez, 2012), arbitrary diagnostic thresholds (Widiger, 2001), heterogeneity within diagnostic categories (Verheul, Bartak, & Widiger, 2007; Westen & Arkowitz-Westen, 1998), and poor operationalization of the general criteria (Bornstein, Bianucci, Fishman, & Biars, 2014; Morey, Benson, Busch, & Skodol, 2015; Skodol, 2012). To address these shortcomings, a new model for personality pathology assessment and diagnosis was introduced in a separate Section III of the fifth edition of the *DSM* (*DSM–5;* American Psychiatric Association, 2013). This section, entitled "The Alternative Model for Personality Disorders (AMPD)," was described as an alternative diagnostic system requiring further research. At the time of *DSM–5* publication, there was limited evidence supporting the reliability, validity, and clinical utility of the AMPD (Clarkin & Huprich, 2011; Shedler et al., 2010; Skodol, Morey, Bender, & Oldham, 2013; Zimmerman, 2013). However, its inclusion in Section III stimulated much needed research (Krueger, Hopwood, Wright, & Markon, 2014; Morey et al., 2015).

The alternative model for personality disorders

The AMPD is based on a two-step approach for PD evaluation, introducing an assessment of personality functioning and a personality trait evaluation. The first step

CONTACT Tore Buer Christensen Tore.buer.christensen@sshf.no Psykiatrisk avdeling Arendal, Sørlandet Sykehus, boks 783 Stoa, 4809 Arendal, Norway.
*Author is also currently based at Unit for Group Psychotherapy, Jessheim District Psychiatric Center, Akershus University Hospital, Lørenskog, Norway.

(termed Criterion A) is intended to capture the severity of impairment in personality functioning, whereas the second step (Criterion B) characterizes personality traits or style (Bender, Skodol, First, & Oldham, 2018; Hopwood et al., 2011). A PD diagnosis requires both sufficient impairment and at least one pathological personality trait. Criterion A is evaluated based on a continuous rather than categorical scale, using the Level of Personality Functioning Scale (LPFS). The LPFS assesses impairments of 12 subdomains of personality functioning (listed in Table 1), summarized within four main domains: identity, self-direction, empathy, and intimacy. LPFS scores range from 0 (*little or no impairment*) to 4 (*extreme impairment*). LPFS Level 2 (moderate impairment) is set as the threshold for PD diagnosis, as this reportedly maximizes the sensitivity and specificity of PD identification (Morey, Bender, & Skodol, 2013). In the study by Morey, every level for the four domains was rated dichotomously (present–not present). However, when using tailored instruments for the assessment of LPFS, the 12 subdomains are rated separately. To obtain a final LPFS score, the sum scores of the 12 subdomains are averaged. For instance, a patient with a sum score of 18 will have a final LPFS level of 1.5, and a patient with a sum score of 20 will have a final LPFS level of 1.67. This calls for a clarification of how to define Level 2. Should the cutoff be placed at 1.5, at 2.0, or somewhere in between? We propose to define Level 2 as situating within the interval from 1.5 to 2.5. However, there is no empirical foundation for this definition yet.

The LPFS is considered a major advancement by many (Clarkin & Huprich, 2011; Livesley, 2012), representing acknowledgment of the need to assess both the presence and severity of personality pathology, rather than merely a specific PD category (Bornstein & Huprich, 2011; Tyrer, Crawford, & Mulder, 2011). It has further been suggested that the LPFS might serve in screening for PD diagnosis (Skodol, Morey, Bender, & Oldham, 2015), facilitating the planning of more individualized treatment, and tracking impairment over time. However, concerns have been raised related to the perceived complexity of the assessment, which might require extensive training of the individual administering the scale (Zimmermann et al., 2012). It has also been argued that the LPFS might be too theory-laden (Pincus, 2011) and could be improved by the use of more "neutral" language familiar to most clinicians (Pilkonis, Hallquist, Morse, & Stepp, 2011). Nonetheless, publication of the AMPD has enabled clinical researchers from interpersonal, psychodynamic, cognitive-behavioral, and social cognitive theoretical perspectives to map the LPFS onto their core theoretical constructs (DeFife, Goldberg, & Westen, 2015; Mancke, Herpertz, & Bertsch, 2015; Pincus, 2011; Waugh et al., 2017; Zimmermann et al., 2012).

Instruments specifically developed for LPFS assessment

Although existing reliable instruments were important in the development of the LPFS (Bender, Morey, & Skodol, 2011), the AMPD was introduced without any measurement tailored for assessment of the scale. Currently, three self-rating instruments are available to assess the LPFS (Bach & Hutsebaut, 2018; Huprich et al., 2017; Hutsebaut, Feenstra, & Kamphuis, 2015; Morey, 2017), as well as three clinician-rated instruments: the Clinical Assessment of the Level of Personality Functioning Scale (CALF; Thylstrup et al., 2016), the Semi-Structured Interview for Personality Functioning DSM–5 (STiP–5.1; Berghuis, Hutsebaut, Kaasenbrood, de Saeger, & Ingenhoven, 2013), and the Structured Clinical Interview for the DSM–5 Alternative Model for Personality Disorders, Module I (SCID–5–AMPD I; Bender et al., 2018).

The CALF includes structured questions concerning each of the four domains of the LPFS, and emphasizes the underlying processes between the rater and the interviewee. In a pilot study, interrater reliability (IRR) estimates were examined in a study using data of 36 patients (12 having a PD; Thylstrup et al., 2016). Each CALF interview was conducted by one of six experts and was video recorded. The recorded interviews were then corated for LPFS by two of the experts who did not conduct the interview. The six experts were trained to perform CALF interviews, but not to evaluate LPFS. ICC was estimated to be .31 to .58 for the domains,

Table 1. ICC video-based interrater reliability.

	ICC	95% CI
Identity	.94	[.88, .98]
Sense of self	.94	[.87, .94]
Self-esteem	.84	[.70, .93]
Emotional range and regulation	.83	[.69, .93]
Self-direction	.94	[.87, .98]
Ability to pursue meaningful goals	.87	[.76, .95]
Constructive, prosocial standards	.84	[.70, .94]
Self-reflective functioning	.91	[.82, .96]
Empathy	.90	[.80, .96]
Understanding and appreciation of others' experiences and motivations	.81	[.66, .92]
Tolerance of differing perspectives	.84	[.70, .93]
Understanding of effects of own behavior on others	.83	[.70, .93]
Intimacy	.89	[.80, .96]
Depth and duration of connections	.83	[.70, .93]
Desire and capacity for closeness	.77	[.60, .90]
Mutuality of regard reflected in behavior	.77	[.60, .90]
Total LPFS	.96	[.92, .98]

Note. CI = confidence interval; ICC = intraclass correlation coefficient, two-way random, single, absolute agreement; LPFS = Level of Personality Functioning Scale.

and .54 for the total LPFS. The authors concluded that the calculated IRR was too weak to consider this instrument as a stand-alone assessment.

The STiP–5.1 (Berghuis et al., 2013) is used to assess all 12 subdomains (see Table 1) of personality functioning. This instrument has a funnel structure, meaning that the interviewer starts with an open question and then narrows down possible levels based on the given response. If this response does not provide sufficient information for scoring, help questions can be asked. If it remains unclear which level is to be scored, dichotomous test questions are provided to make a final determination of the patient's level. The aim of using such a structure is to be time-efficient and to increase parsimony and clinical utility. In a recent study examining the reliability of this instrument, 12 regular staff psychologists with different levels of training and experience examined a clinical sample of 40 treatment-seeking participants (80% with a PD) and 12 relatives (Hutsebaut, Kamphuis, Feenstra, Weekers, & De Saeger, 2016). The interviews were video-recorded and then independently scored by one of the authors. ICC values ranged from .64 to .80 for the subdomains, and the ICC was .71 for the total LPFS.

The SCID–5–AMPD I is described later in this article.

Other investigations of LPFS assessment

Other studies have investigated LPFS assessment based on clinical interview information without using a specifically tailored instrument. In a study applying a video-recording design in a sample of 109 outpatients, Few et al. (2013) examined the IRR of the LPFS assessments based on information from the Structured Clinical Interview for Axis II Disorders (SCID–II; First, Spitzer, Gibbon, Williams, & Benjamin, 1994). IRR was poor, with ICC ranging from .47 to .49 for the domains, possibly because the SCID–II does not yield all of the information necessary for LPFS determination. Zimmermann et al. (2014) focused on how reliably psychology students could assess the LPFS, investigating the IRR of LPFS assessments. In their study, 10 female inpatients (50% with one or more PD diagnoses) were assessed by 22 untrained, clinically inexperienced students. The students used an adapted, multi-item version of the LPFS to score impairment in functioning based on video recordings of experts performing interviews following the guidelines of the Operationalized Psychodynamic Diagnosis system (Force, 2008). ICC estimates ranged from .25 to .63 for the four main domains, and the ICC was .51 for the total LPFS. However, the reliability of the LPFS mean score across raters was extremely high (ICC = .96). In another recently published study (Garcia et al., 2018), 13 advanced clinical psychology doctoral students with minimal familiarity with the LPFS evaluated clinical vignettes though three sessions of learning. The estimates of reliability (ICC) improved for each learning session, resulting in ratings for the domains in the range of .59 to .75, for global LPFS rating .81. This supports the conclusion of the study by Zimmermann et al., indicating that some criticisms related to the complexity of the LPFS might have been premature.

The SCID–5–AMPD I

The developers of the LPFS later introduced the SCID–5–AMPD I. Like the STiP–5.1, this instrument has a funnel structure, starting with open questions for each domain to obtain a general impression of the level of personality functioning. However, in contrast to the STiP–5.1, the SCID–5–AMPD I includes specific follow-up questions for each LPFS level (0–4) and for each subdomain. Moreover, whereas follow-up questions are optional in the STiP–5.1, the SCID–5–AMPD I requires that the interviewer always poses a number of follow-up questions. Thus, the SCID–5–AMPD I is more complex and might take more time to complete. These properties could be perceived as limitations; however, they can also be advantageous, potentially resulting in richer information, allowing for more precise measurement, and thus increasing the likelihood of obtaining good IRR. To date, no studies have examined the IRR of the SCID–5–AMPD I.

Limitations of studies of interrater reliability

The importance of using a test–retest design when examining reliability is emphasized in many studies (e.g., Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Helzer et al., 1977), including in several papers describing the *DSM–5* field trials (e.g., Kraemer, Kupfer, Narrow, Clarke, & Regier, 2010). Test–retest IRR field trials of the LPFS were planned, but not conducted (Zachar, Krueger, & Kendler, 2016). Although most reliability studies, including the studies of LPFS, use a video-recording design, such investigations cannot measure variation in the patient's history and self-presentation, or differences in how the interview is performed. Test–retest analyses are more rigorous, providing data from two different interview situations. Chmielewski, Clark, Bagby, and Watson (2015) compared estimates of the diagnostic reliability of *DSM–IV* (defined as the extent to which the patient would receive the same diagnosis by two different raters) using a video-recording design compared to a test–retest design, and demonstrated superior estimates with the video-recording design. However, there are also some limitations with a test–retest design. Most important is the variability due to the interval between the two interview sessions. Although it is unlikely that a patient's personality function will truly change during an interval of less than 2 months (Verheul et al., 2008), differences in the patient's story between interviews can also contribute to variance. Differences in interview style of the different raters is a third source of variance with this method, but could also be considered a strength because this will mimic most clinical situations and give a more realistic estimate. However, greater information about reliability can be acquired when both methods are applied (Grove et al., 1981; Helzer et al., 1977).

This study

This study is a part of the larger Norwegian AMPD Multisite Project (NorAMP) research project and aims at evaluating the interrater reliability of the LPFS as assessed by the Norwegian translation of the SCID–5–AMPD I. Reliability was tested using two different designs: a video-recording design and a short-interval test–retest design (see Weertman, Arntz, Dreessen, van Velzen, & Vertommen, 2003).

Methods

Participants and recruitment sites

The Norwegian AMPD multisite project

NorAMP is a multisite research project aiming to evaluate the reliability, clinical utility, and validity of the AMPD. A total of 286 patients were recruited from different levels of psychiatric care within six hospitals in Norway between March 2015 and March 2017. Participants were recruited to the study by their therapists, and recruitment sites included general mental health inpatient and outpatient departments, group psychotherapy outpatient and day treatment units, two substance abuse units (both outpatient and inpatient), and a prison clinic. The group psychotherapy units were all part of the Norwegian Network of Personality-Focused Treatment Programs (Karterud & Wilberg, 2007), a large collaborative network of clinical units specializing in PD assessment and treatment. Exclusion criteria were as follows: schizophrenia spectrum disorder (except schizotypal PD), sequelae after brain injury, pervasive developmental disorders (i.e., autism spectrum disorders), mental retardation, severe ongoing substance abuse, and lack of understanding of Norwegian language.

To ensure a broad range of personality functioning in the total sample, a group of 35 nonclinical participants who were not undergoing clinical treatment during the last 5 years was also included. These were recruited by an information poster among students and employees at the University of Agder, University of Oslo, and Sorlandet Sykehus.

The NorAMP sample

The final study sample included 317 participants, all recruited between March 2015 and March 2017. In the total sample, including the nonclinical sample ($n = 35$), 207 (65%) were female, the mean age was 32.3 years (range = 16–72), and average level of education was 4.4 years of school after secondary school ($SD = 2.8$). The mean number of SCID–II criteria was 11.1 ($SD = 8.1$; range = 0–49) and mean number of PD diagnoses was 1.05 ($SD = 1.1$; range = 0–7). For the 192 (61%) who fulfilled criteria for a *DSM–IV* PD diagnosis, these were distributed as follows: 81 with avoidant PD (APD; 42%), 70 with borderline PD (BPD; 37%), 44 with PD not otherwise specified (PD NOS; 23%), 30 with antisocial PD, and 30 with paranoid PD (16%), 21 with obsessive–compulsive PD (11%), and 14 with dependent PD (7%). In the clinical sample, 83.7% had one or more Axis I diagnosis ($M = 1.6$, $SD = 1.4$; missing 2.5%, $n = 7$). The most frequent were major depression (27%), social phobia (19%), posttraumatic stress disorder (PTSD; 13%), substance abuse (12%), generalized anxiety disorder and dysthymia (both 10%), and panic disorder with agoraphobia (9%). For 74 participants (23%), including the nonclinical sample, no information concerning Axis I diagnosis was provided.

IRR subsample I

For the first 85 participants included in the NorAMP study, the interviews were video-recorded. These were grouped according to rated global level where after 17 videos were randomly selected, using a Web-based research randomizer (Urbaniak & Plous, 1997). In this subsample, 11 (65%) were female, mean age was 31.6 (range = 19–59), and average level of education was 4.6 years after secondary school ($SD = 3.1$). Mean number of SCID–II criteria was 11.4 ($SD = 11.2$, range =0–30), and mean number of PD diagnoses was 1.3 ($SD = 1.3$, range = 0–4). Nine participants fulfilled criteria for a PD diagnosis (53%), which were distributed as follows: eight had APD, four had BPD, two had antisocial PD, two paranoid PD, and one had dependent PD. In this sample, 76.5% had an Axis I diagnosis ($M = 1.9$, $SD = 1.1$, range = 1–4), with major depressive disorder (58.9%) as the most frequent, followed by panic disorder (12.0%), PTSD (12.0%), drug abuse (12.0%), social phobia (12.0%), agoraphobia (12.0%), and attention deficit disorder (12.0%). For three participants, no information regarding Axis I diagnosis was provided (drawn from the nonclinical sample).

Participants in the video-based IRR study were not significantly different from the main sample with respect to age ($t = .29$, $p = .766$), sex ($\chi^2 = .003$, $p = 1.0$), number of PD criteria ($t = .231$, $p = .818$), and mean LPFS level ($t = 1.00$, $p = .317$).

IRR subsample II

From January to July 2016, 34 patients participated in the test–retest IRR study. Due to practical considerations, all participants were recruited from the Oslo region. One patient was excluded due to autism spectrum disorder, diagnosed after study inclusion. All diagnoses were based on the use of semistructured interviews (see later). Among the remaining 33 participants, 24 (73%) were female, mean age was 29 years (range = 20–55 years), and mean level of education was 4.3 years after secondary school ($SD = 2.7$). The mean number of SCID–II criteria was 9.8 ($SD = 5.1$), mean number of PD diagnoses was 0.91 ($SD = 0.38$, range = 0–3). Among the 28 participants receiving a PD diagnosis, 13 had a BPD (46%), 8 a PD NOS (29%), and 6 received a diagnosis of APD (21%). A total of 28 patients (85%) received an Axis I diagnosis ($M = 1.6$, $SD = 1.2$, range = 1–5). The most common Axis I diagnoses were recurrent depression (42%), social phobia (24%), generalized anxiety disorder (15%), dysthymia (12%), and PTSD (12%).

Patients participating in the test–retest study were significantly different from the main clinical sample for age (32.9 vs. 28.9, $t = 2.1$, $p = .033$), but not for sex ($\chi^2 = .059$, $p = .438$), mean level of personality functioning ($t = 1.3$,

$p = .191$), and number of PD criteria ($t = 1.04$, $p = .299$). Because only patients were included in the test–retest study, nonclinical controls were not included in the latter comparisons.

Diagnostic assessments

Mini international neuropsychiatric interview

Symptom disorders were assessed by experienced referring therapists who used the Mini International Neuropsychiatric Inventory (MINI), a short structured diagnostic interview for *DSM–IV* and International Classification of Diseases (10th ed. [ICD–10]) psychiatric disorders. Reliability and validity of the MINI are both considered to be good (Sheehan et al., 1998). In this study we used the Norwegian version 5.0.0. of the MINI, revised in 2007. The interviews were conducted by the referring therapists ($M = 14$ years of experience, $SD = 10$), including 44.5% psychologists, 27.8% psychiatrists, 19.6% social workers or nurses and 8.1% with another degree. IRR was not tested.

Structured clinical interview for Axis II disorders

Prior to inclusion, referring therapists also performed the SCID–II, a semistructured interview used to assess the 10 *DSM–IV* PDs and PD NOS (First et al., 1994), which showed good interrater and test–retest reliability in PD samples (Weertman et al., 2003). Referring therapists were trained by the National Knowledge Center for Personality Disorders at the Oslo University Hospital. The quality of the SCID–II assessments was ascertained by consensus training of all referring therapists, using video-recorded interviews. During both the initial training and the video sessions, independent ratings and discrepancies were discussed. The reliability of the SCID–II diagnoses was not evaluated in this sample. However, a former study from the Norwegian Network of Personality-Focused Treatment Programs reported kappa coefficients of the three PDs: APD ($\kappa = .75$), BPD ($\kappa = .66$), and paranoid PD ($\kappa = .71$). This indicates acceptable diagnostic reliability within the network, from where most patients in the test–retest study (31 of 33) were recruited.

Iowa Personality Disorder screen

The Iowa Personality Disorder Screen is an 11-item screening instrument for the presence of PD, which was used in the recruitment procedure for nonclinical participants to exclude PDs. Sensitivity and specificity estimates in psychiatric samples have been high (Langbehn et al., 1999). The items correspond to diagnostic criteria for PDs and are rated dichotomously (yes or no).

The SCID–5–AMPD I

The SCID–5–AMPD I is a semistructured interview that covers the 12 subdomains of the LPFS (Bender et al., 2018). The instrument starts with eight general overview questions addressing how the subject relates to himself or herself and to others. For each of the 12 subdomains, the assessment begins with screening questions. For example, for the subdomain of identity, sense of self, the first question is "Do you sometimes have the experience of not really knowing who you are or how you are unique in the world?" Based on clinical judgment and screening, the rater is instructed to ask questions for each subdomain corresponding to the level at which the interviewee might be functioning. There are one to six specific questions for each level—for example, the questions for Level 2 of the subdomain sense of self are "Do you depend on other people's opinions in order to know who you really are?" and "Is it hard for you to know who you are without knowing what other people think of you?" The rater explores increasing levels of impairment until the interviewee clearly does not qualify for that level. The text also includes the descriptions of all levels, which were taken directly from the LPFS, for use as anchor points for the rating. For each subdomain a rating of level is set, giving three scores for each domain, resulting in an average score for each domain. When all 12 subdomains are rated, a global LPFS score is set as an average of all four domain scores. In our study, the interview included a few questions concerning demographics and former psychopathology, and the interviewer had access to the original referral, providing variable brief background information about the patient. Raters were instructed to mark the text to indicate which levels they explored. For use in our study, both the LPFS and the SCID–5–AMPD I were translated into Norwegian by members of the Department of Personality Psychiatry, Oslo University Hospital. No back-translation procedure was performed.

Raters and training

Among the seven raters in the test–retest IRR study, three were experienced clinicians, including two clinical psychologists and one psychiatrist (all male). They underwent training in LPFS assessment during a two-day workshop by Dr. Donna Bender, along with the other raters in the NorAMP study (seven experienced clinicians altogether). The four inexperienced clinicians in the test–retest IRR study included three undergraduate psychology students and one undergraduate medical student (all female) who were trained by two of the experienced raters several weeks before inclusion in the test–retest study. The content and duration of their training was practically identical to that in the workshop provided by Dr. Bender. The training included an introduction to the instrument, the use of nine written case vignettes, one demonstration of assessment by a role-play, and one video interview. Through the training, global LPFS scores for the written vignettes were set by each rater independently, and then discussed in plenum. For both the demonstration and the video, this also included scores of domains and subdomains. The procedure was repeated until consensus was achieved.

Diagnostic procedure

All patients were referred by therapists in mental health service units (referring therapists). The raters conducting the

SCID–5–AMPD I had no access to the results of SCID–II assessments. There was a maximum interval of 5 weeks between the SCID–II and performance of the SCID–5–AMPD I.

The sample of participants not undergoing clinical treatment was screened for PD using the Iowa Personality Disorder Screen (Langbehn et al., 1999), which was administered over the phone by an experienced rater. These participants were recruited by an informational poster that invited them to participate by calling a telephone number to receive further information about the study.

Video-based IRR study

The 17 participants were drawn from a pool of 85 video-recorded assessments of SCID–5–AMPD I, conducted by one of seven experienced raters. The number of recordings from each rater ranged from two to six, and each selected interview was scored by the four remaining raters independently. Following this procedure, therefore, each participant was given five independent scores.

Test–retest IRR study

In the test–retest study, the SCID–5–AMPD I was administered by seven raters: three experienced clinicians and four inexperienced clinicians. All 33 patients were assessed separately by two raters, performing the interviews with a maximum interval of 2 weeks. We attempted to assign raters according to a rotation schedule, with the aim of pairing each rater with all other raters at least once. The schedule was also intended to balance the raters in the first and second interview positions, and included all possible combinations of rater pairs (7 raters squared $\times$ 2 rater positions = 98 possible combinations). However, due to rater availability, it was not possible to perfectly execute the rotation schedule to include all possible combinations. Thus, some raters were paired together more often than others. The combination regarding experience was as follows: 15 patients were assessed by two inexperienced raters, 13 by one experienced and one inexperienced rater, and five participants were assessed by two experienced raters. Raters were blinded to the other rater's evaluations and to the SCID–II results.

The mean duration of the interval between the two ratings was 9.2 days ($SD = 5.4$ days). Each patient received a 500 NOK (approximately $50) gift card for being interviewed a second time. All participants completed both interviews. The interviews lasted an average of 80 min, with the mean duration being 72 min for experienced clinicians and 83 min for inexperienced clinicians. Notably, this time period included questions concerning demographic data and former psychopathology, which are now officially included as a part of the SCID–5–AMPD I.

Statistics

Our aim was to estimate IRR using two methods: a video-based approach and a short-interval test–retest approach. Generalizability to other raters was important with both protocols. For both methods, we acquired the same number of ratings for every rated participant, and the assumptions of normality distribution and absence of outliers were met. Thus, IRR was evaluated by means of ICC calculated for global LPFS, domains, and subdomains. The domain scores were calculated as the average of each of the three subdomain scores, and the global LPFS score as the average of the four domain scores. In the test–retest study, we also used Cohen's kappa to assess the level of agreement between raters with regard to dichotomous ratings based on the LPFS (PD vs. no PD), and used t tests for group comparisons. Different formulas are used to calculate ICC depending on the study design. We provide information about the model, form, and type of ICC calculated, according to the categorization system of Shrout and Fleiss (1979) and McGraw and Wong (1996; see also Trevethan, 2017).

In the video-based IRR study, 17 participants were randomly drawn from a group of 85 participants, and five raters were randomly drawn from a group of seven raters. Each rater assessed all 17 participants. We used a two-way random effect model to estimate the extent to which the five raters gave the 17 participants similar personality functioning scores using the SCID–5–AMPD 1 instrument. In this model, the variation in the data is regarded as coming from two sources: the participants and the raters. Regarding forms of ICC, there are two choices: single measures and average measures. In our study, it was not of interest to average the ratings of each rater prior to ICC analysis—rather, we wanted to estimate the reliability of each single rater's score. Thus, we use a two-way random effect model, single measure, typically expressed as ICC (2,1). There are two types of each combination of model and form, referred to as consistency and absolute agreement. Here, we wanted to estimate how similar the five raters' scores were, not only how they were correlated; therefore, we applied the absolute agreement approach.

The design of the second test–retest reliability IRR study was not fully crossed, as each participant was rated by two different pairs of raters who were considered randomly selected. Therefore, we applied a one-way random effect model for the ICC analyses, using the same form and type of ICC as in the video-based study (ICC (1,1))

We deemed it reasonable to interpret the results of the video-based IRR study following the guidelines of Cicchetti (1994). According to these, coefficients below .40 indicate poor interrater reliability, those between .40 and .60 are indicative of fair agreement, those between .60 and .74 are considered good, and coefficients higher than .75 are regarded as excellent. For the test–retest IRR study, we report the criteria used to examine the reliability in *DSM–5* field trials based on a test–retest design with two different raters (Chmielewski et al., 2015; Clarke et al., 2013; Kraemer, Kupfer, Narrow, Clarke, & Regier, 2010). The ranges are as follows: excellent ($>$ .80), very good (.60–.79), good (.40–.59), questionable (.20–.39), and unacceptable ($<$ .20). For the test–retest field trials, the *DSM–5* Taskforce regarded ICC values of $>$ .40 as acceptable, with a maximum range of 0.5 for the 95% confidence

interval (Kraemer et al., 2012). All statistical analyses were performed using IBM SPSS Statistics 23.0 (IBM Corporation, Armonk, NY).

In our study, LPFS was calculated as follows: Level 0, mean LPFS 0–.49; Level 1, .50–1.49; Level 2, 1.50–2.49; Level 3, 2.50–3.49; and Level 4, 3.50–4.00.

Ethics

All participants gave written consent before participating in this study. The project was approved by the Regional Committee for Medical and Health Research Ethics.

Results

The video-based IRR study

The ICC for the overall LPFS scoring was .96 (95% CI [.92, .98]). The ICC values for the four domains were somewhat smaller, but still large, ranging from .89 (intimacy) to .95 (identity and self-direction). The smallest ICC value was .77 for the two identity subdomains mutuality of regard reflected in behavior and desire and capacity for closeness. The other ICC values were all > .80 (Table 1).

The test–retest IRR study

The ICC for the total LPFS score was .75 (95% CI [.55, .87]; Table 2). ICC values for the four LPFS domains ranged from .59 (identity) to .80 (self-direction). The smallest ICC values were .32 for the subdomain self-esteem and .24 for the subdomain mutuality of regard reflected in behavior. A paired-sample t test indicated that LPFS scores did not significantly differ between first interview (M LPFS = 2.06, SD = .67) and the second interview (M LPFS = 2.05, SD = .92), t = .15, p = .82. Three patients represented outliers (Patients 14, 19, and 33) in Figure 1. Inspection of one of these pairs of interviews revealed that the subdomain-specific screening questions were interpreted very differently for these patients. The first rater asked questions related to Levels 0 to 1 and concluded with a score of Level 0 (no impairment), whereas the second rater asked questions related to Levels 2 to 4 and concluded by setting a score of 4 (extreme impairment). This pattern was apparent in the ratings of Patients 14, 19, and 33. Closer inspection revealed that all ratings for these three patients were performed by inexperienced raters. Correlation analyses excluding these patients resulted in a substantial increase of the ICC estimates, improving from .75 (CI [.55, .87]) to .88 (CI [.77, .94]) for the mean LPFS; for the domains the range of estimates increased from .59 through .80 to .65 through .87. We further used Cohen's kappa coefficient to explore the degree of agreement related to LPFS threshold levels for a PD diagnosis. Using scores averaged over the 12 subdomains, the diagnostic threshold for a PD can be determined in two ways: by applying a cutoff score of 1.5 or by applying a cutoff score of 2.0. As far as we know, how to define the threshold for Level 2 is not clarified, and both were calculated. For the first cutoff score (i.e., a mean LPFS threshold level of 1.5) the Cohen's kappa was .57. Using a mean LPFS threshold level of 2 for the PD cutoff, we found a substantially larger Cohen's kappa of .70.

Discussion

This study is the first to investigate reliability of the SCID–5–AMPD I, an instrument that is specifically tailored for assessing levels of personality functioning as presented in Section III of *DSM–5*. Among several studies focusing on LPFS assessment, this study is the first to measure IRR using a test–retest design.

Main findings

High IRR estimates were obtained by co-ratings of video-recorded SCID–5–AMPD I assessments. All estimates were in the excellent range, both for domains and subdomains.

Unsurprisingly, IRR evaluated using a test–retest setup was lower than that estimated in the video-recorded test design. However, the estimates from the test–retest setup

Table 2. ICC test–retest-based interrater reliability.

	ICC	95% CI
Identity	.59	[.32, .77]
Sense of self	.59	[.32, .78]
Self-esteem	.32	[−.02, .59]
Emotional range and regulation	.40	[.07, .65]
Self-direction	.80	[.63, .89]
Ability to pursue meaningful goals	.71	[.50, .85]
Constructive, prosocial standards	.60	[.34, .78]
Self-reflective functioning	.72	[.50, 85]
Empathy	.69	[.47, .84]
Understanding and appreciation of others' experiences and motivations	.53	[.24, .74]
Tolerance of differing perspectives	.66	[.42, .82]
Understanding of effects of own behavior on others	.66	[.41, .81]
Intimacy	.63	[.37, .80]
Depth and duration of connections	.60	[.33, .78]
Desire and capacity for closeness	.57	[.30, .76]
Mutuality of regard reflected in behavior	.24	[−.10, .53]
Total LPFS	.75	[.55, .87]

Note. CI = confidence interval; ICC = intraclass correlation coefficient, one-way random, single, absolute agreement; LPFS = Level of Personality Functioning Scale.

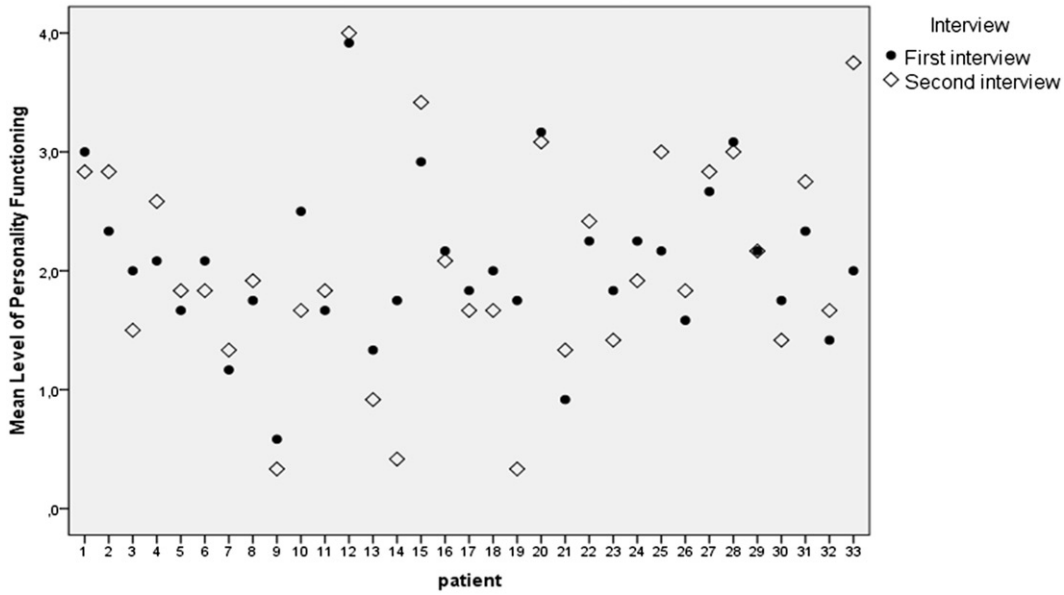

Figure 1. Level of Personality Functioning Scale (LPFS) ratings in the test–retest study ($n = 33$).

were within the good to very good range for the domains and overall LPFS score, and were still high compared to previously published studies using a video-based design. Finally, when dichotomous LPFS scores were evaluated in relation to a diagnostic threshold for PD, the kappa scores were good when applying an LPFS threshold of 2 for a PD diagnosis, but were moderate when using an LPFS threshold of 1.5.

High IRR with a video-based design

In our study, the video-based design resulted in a higher IRR than that reported for a comparable instrument, the STiP–5.1 (ICC for domains ranging from .89–.94 vs. .64–.80). Moreover, both of these specifically designed instruments achieved a higher IRR than the values reported in studies using instruments designed for other purposes, such as the SCID–II and Operationalized Psychodynamic Diagnosis system–Level of Structural Integration Axis (OPD-LSIA) (Di Pierro, Benzi, Madeddu, & Preti, 2016; Few et al., 2013; Zimmermann et al., 2014). It is reasonable to expect that the use of an instrument specifically designed to assess LPFS would contribute to enhanced reliability.

However, the results of our video-based IRR study should be interpreted cautiously. When using an instrument with a funnel structure, like the STiP–5.1 and SCID–5–AMPD I, only selected levels are examined based on the interviewee's responses to screening questions. This can inflate the IRR estimates in two ways. First, the information provided to the second rater (the observer) will rely on the judgment of the first rater (the interviewer). Second, this structure can enable the second rater to determine the conclusions of the first rater. Nevertheless, video-based test designs provide advantageous opportunities for simultaneous assessments. In a video-based study, no variance is introduced due to the time interval (during which the interviewee's clinical status could change) or due to possible effects of patients being tested twice (e.g., an interviewee trying to remember his or her response from the first interview or minimizing responses to shorten the interview).

Good to very good IRR with a test–retest design

In our test–retest study, ICC estimates ranged from .59 to .80 for the four domains, and the ICC value was .75 for the total LPFS. These values indicate a remarkably high IRR for a test–retest design. According to the criteria used in the *DSM–5* field trials, three of the four main domains were within the very good range, with self-direction achieving the highest ICC estimate of .80. This domain refers to the more concrete and recognizable aspects of functioning, such as pursuit of life goals and accomplishments, and might thus be easier to assess.

The least reliably assessed domain in the test–retest IRR study was identity (ICC = .59). Notably, the subdomains of self-esteem and emotional range and regulation had low ICC estimates of .32 and .40, respectively. One possible explanation is that this domain is difficult to consistently assess due to fluctuations in the patient's self-states, which could be especially relevant in our study because our sample was characterized by a high prevalence of BPD patients, who by definition have problems with identity (Wilkinson-Ryan & Westen, 2000). The identity domain is also arguably an abstract or ambiguous concept that is difficult to measure with a high degree of precision. In line with this, when reporting self-relevant information, participants are also likely to report what comes to mind easily at the point of time of the assessment, which could have a negative impact on reliability (Oyserman, 2001). This resonates with the standpoint of the ICD–11 workgroup for PDs, writing that "an accurate assessment of self-pathology of personality is highly complex and beyond the expectations of most practitioners" (Tyrer, Reed, & Crawford, 2015, p. 723). However, in the *DSM–5* field trials of "cross-cutting"

dimensional measures (Narrow et al., 2013), the self-report item "Not knowing who you really are or what you want in life" had a test–retest ICC of .66, which was only slightly lower than the estimated ICC of .68 for an item addressing interpersonal problems, "Not feeling close to other people or enjoying your relationships with them." These results suggest that self-functioning and interpersonal functioning can be measured with equal reliability. With regard to this study, it is notable that the third identity subdomain, sense of self, had an ICC estimate within the good range. Thus, we believe that it is premature to assert that self-pathology is too complex to be reliably assessed in clinical practice. Rather, there might be a need to reevaluate how the questions are formulated.

Within the intimacy domain, the subdomain of mutual regard reflected in behavior showed a low ICC value of .24. Some patients might experience the topic of this subdomain as a question of moral standard, which could have two main implications. First, some questions might be considered confrontational. For example, the following questions were asked in relation to Level 2: "Do you primarily choose situations or relationships that clearly benefit you in some way, help you get ahead, or reflect well on you?" and "Do you generally only do things with or for others if there is something in it for you?" Interestingly, scores corresponding to this level were assigned in only 5 of the 66 ratings in our sample, possibly explaining the low estimate of reliability. The second implication is that the low estimates could be related to an artifact termed the *social desirability* hypothesis of the retest effect (Jorm, Duncan-Jones, & Scott, 1989). This effect is confined to questions assessing negative self-characteristics, in which interviewees tend to give more favorable answers on retesting (Durham et al., 2002). However, the mean scores for this subdomain did not significantly differ between the two interviews, indicating that this effect was not present in our study.

Notably, similar concerns were raised in another European study of LPFS (Zimmermann et al., 2014). The authors suggested future adjustments to the questioning style, to include less direct questions and greater focus on reflective functioning (Fonagy, Target, Steele, & Steele, 1998). In the AMPD, the threshold for a PD diagnosis is a moderate or more severe impairment in personality function, indicated by LPFS Level 2 or higher. This threshold was included in the manual based on a study by Morey and colleagues, which demonstrated that an LPFS level of 2 (moderate impairment) distinguished PDs from no PD or other mental disorders with maximum combined sensitivity and specificity, both in general and for each of the six specific PDs included in the AMPD (Morey et al., 2013). Our present analyses revealed that the agreement was better when the threshold for moderate impairment was set as +2.0 rather than at +1.5, which was used in a recent study (Morey, 2017).

Complexity of assessment and training of raters

Our results showed no statistically significant difference in IRR according to level of experience. Several authors anticipated that reliable LPFS assessment would likely require extensive training (Pilkonis et al., 2011; Pincus, 2011; Zimmermann et al., 2012). The use of a similar well-established instrument, the OPD–LSIA, requires 60 hr of standard training (Zimmermann et al., 2012). Other studies of LPFS assessment have reported varying degrees of rater training, ranging from no training at all (Zimmermann et al., 2014) to extensive training (Hutsebaut et al., 2016). Relative to other studies, the 2 days of training administered in our study is in the midrange and is feasible to implement in most clinical settings.

Another concern that has been raised regarding LPFS reliability is the need for sufficient guidelines (Pincus, 2011). An extensive manual has been developed for the STiP–5.1 instrument (Berghuis et al., 2013). Although a User's Guide for the three modules of the SCID–5–AMPD has now been developed, at the time of our study, no scoring manual was available for the SCID–5–AMPD I. The raters in this study received a simple one-page set of instructions, including an explanation of the funnel structure. This is now included as an integrated part of the introduction of the official version of Module I. Overall, our results indicated that the SCID–5–AMPD I can be reliably used to assess LPFS with an amount of training that should be reasonable for most clinical settings.

Limitations and future directions

Our results should be interpreted cautiously. As discussed, there are several limitations related to the video-based method. Additionally, the samples in our study were small. In the test–retest IRR study, the precision for three of the subdomains was poorer than is considered acceptable in the *DSM* field trials, with a lower limit of 0.5 for the 95% CI (Clarke et al., 2013). Analysis of a larger sample might have resulted in a smaller CI, and thus increased the precision of our estimates. However, the precision was acceptable for the four domains.

There might also be limitations related to the representativeness of our sample. Within the sample included in our test–retest IRR study, up to 86% were diagnosed with a PD, which is a considerably higher frequency than found in an ordinary outpatient population. Furthermore, most patients had been referred for long-term psychotherapy. However, the focus of our study was to evaluate core features among people with PD. Although our sample did not cover the full range of PD diagnostic categories, the distributions of both gender and diagnoses were in line with the findings of previous clinical studies (Karterud et al., 2003; Silberschmidt, Lee, Zanarini, & Schulz, 2015). Considering that the examined instrument was designed for use as a diagnostic procedure when PD is suspected, we expect the findings to be reasonably generalizable. Future studies using a similar method in larger samples are needed to confirm this.

Conclusions

According to our findings, LPFS can be measured with a sufficient degree of reliability using the SCID–5–AMPD

I, even with only a modest amount of training. Further investigations including large samples of representative clinical and community populations are needed to confirm our findings.

Acknowledgments

We are indebted to the patients who volunteered in this study; the staff members at Akershus University Hospital, Lovisenberg Hospital, Namsos Hospital, Oslo University Hospital and Sorlandet Hospital; and the members of the Research Group at Department of Personality Psychiatry and the National Competence Services for Personality Psychiatry.

Funding

The study has been supported by Sorlandet Sykehus and Oslo University Hospital.

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual*, (3rd edn). Washington, D.C.: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington: American Psychiatric Pub.

Bach, B., & Hutsebaut, J. (2018). Level of personality functioning scale-brief form 2.0: Utility in capturing personality problems in psychiatric outpatients and incarcerated addicts. *Journal of Personality Assessment*, 1–11. Advance online publication. doi:10.1080/00223891.2018.1428984

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in DSM-5, part I: A review of theory and methods. *Journal of Personality Assessment*, *93*(4), 332–346. doi:10.1080/00223891.2011.583808

Bender, D. S., Skodol, A., First, M. B., & Oldham, J. (2018). Module I: Structured clinical interview for the level of personality functioning scale. In M. B. First, A., Skodol, D. S., Bender, & J., Oldham *Structured clinical interview for the DSM-5 alternative model for personality disorders (SCID-AMPD)* (Ed.). Arlington, VA: American Psychiatric Association.

Berghuis, H., Hutsebaut, J., Kaasenbrood, A., de Saeger, H., & Ingenhoven, T. (2013). The semi-structured interview for personality functioning DSM-5 (STiP-5). *EESPD Sci News2013* (available at www. esspd. eu/fileadmin/user_upload/ESSPD_Newsletter/Downloads_Newsletter/4-5_ESSPD_Newsletter_January_2014. pdf, accessed December 18, 2014).

Bornstein, R. F., Bianucci, V., Fishman, D. P., & Biars, J. W. (2014). Toward a firmer foundation for DSM-5.1: Domains of impairment in DSM-IV/DSM-5 personality disorders. *Journal of Personality Disorders*, *28*(2), 212–224. doi:10.1521/pedi_2013_27_116

Bornstein, R. F., & Huprich, S. K. (2011). Beyond dysfunction and threshold-based classification: A multidimensional model of personality disorder diagnosis. *Journal of Personality Disorders*, *25*(3), 331–337. doi:10.1521/pedi.2011.25.3.331

Chmielewski, M., Clark, L. A., Bagby, R. M., & Watson, D. (2015). Method matters: Understanding diagnostic reliability in DSM-IV and DSM-5. *Journal of Abnormal Psychology*, *124*(3), 764–769. doi:10.1037/abn0000069

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284. doi:10.1037/1040-3590.6.4.284

Clarke, D. E., Narrow, W. E., Regier, D. A., Kuramoto, S. J., Kupfer, D. J., Kuhl, E. A., … Kraemer, H. C. (2013). DSM-5 field trials in the United States and Canada, Part I: Study design, sampling strategy, implementation, and analytic approaches. *American Journal of Psychiatry*, *170*(1), 43–58. doi:10.1176/appi.ajp.2012.12070998

Clarkin, J. F., & Huprich, S. K. (2011). Do DSM-5 personality disorder proposals meet criteria for clinical utility? *Journal of Personality Disorders*, *25*(2), 192–205. doi:10.1521/pedi.2011.25.2.192

DeFife, J. A., Goldberg, M., & Westen, D. (2015). Dimensional assessment of self- and interpersonal functioning in adolescents: Implications for DSM-5's general definition of personality disorder. *Journal of Personality Disorders*, *29*(2), 248–260. doi:10.1521/pedi_2013_27_085

Di Pierro, R., Benzi, I. M. A., Madeddu, F., & Preti, E. (2016). *Personality pathology assessment: Use of the Level of Personality Functioning Scale by clinically inexperienced raters and associations with the Structured Interview of Personality Organization*. Paper presented at the ESSPD Annual meeting Wien.

Durham, C. J., McGrath, L. D., Burlingame, G. M., Schaalje, G. B., Lambert, M. J., & Davies, D. R. (2002). The effects of repeated administrations on self-report and parent-report scales. *Journal of Psychoeducational Assessment*, *20*(3), 240–257. doi:10.1177/073428290202000302

Few, L. R., Miller, J. D., Rothbaum, A. O., Meller, S., Maples, J., Terry, D. P., … MacKillop, J. (2013). Examination of the Section III DSM-5 diagnostic system for personality disorders in an outpatient clinical sample. *Journal of Abnormal Psychology*, *122*(4), 1057–1069. doi:10.1037/a0034878

First, M. B., Spitzer, R., Gibbon, M., Williams, J., & Benjamin, L. (1994). *Structured clinical interview for DSM-IV axis II personality disorders (SCID-II, version 2.0)*. New York: Biometric Research Department, New York Psychiatric Hospital.

Fonagy, P., Target, M., Steele, H., & Steele, M. (1998). *Reflective-functioning manual, version 5.0, for application to adult attachment interviews*. London: University College London.

Garcia, D. J., Skadberg, R. M., Schmidt, M., Bierma, S., Shorter, R. L., & Waugh, M. H. (2018). It's not that difficult: An interrater reliability study of the DSM–5 Section III alternative model for personality disorders. *Journal of Personality Assessment*, 1–9. Advance online publication. doi:10.1080/00223891.2018.1428982

Grant, B. F., Stinson, F. S., Dawson, D. A., Chou, S. P., & Ruan, W. J. (2005). Co-occurrence of DSM-IV personality disorders in the United States: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. *Comprehensive Psychiatry*, *46*(1), 1–5. doi:10.1016/j.comppsych.2004.07.019

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry*, *38*(4), 408–413.

Helzer, J. E., Robins, L. N., Taibleson, M., Woodruff, R. A., Jr., Reich, T., & Wish, E. D. (1977). Reliability of psychiatric diagnosis. *I. A methodological review. Archives of General Psychiatry*, *34*(2), 129–133. doi:10.1001/archpsyc.1977.01770140019001

Hopwood, C. J., Malone, J. C., Ansell, E. B., Sanislow, C. A., Grilo, C. M., McGlashan, T. H., … Morey, L. C. (2011). Personality assessment in DSM-5: Empirical support for rating severity, style, and traits. *Journal of Personality Disorders*, *25*(3), 305–320. doi:10.1521/pedi.2011.25.3.305

Huprich, S. K., Nelson, S. M., Meehan, K. B., Siefert, C. J., Haggerty, G., Sexton, J., … Baade, L. (2017). Introduction of the DSM-5 Levels of Personality Functioning Questionnaire. *Personality Disorders*. Advance online publication. doi:10.1037/per0000264

Hutsebaut, J., Feenstra, D. J., & Kamphuis, J. H. (2015). Development and preliminary psychometric evaluation of a brief self-report questionnaire for the assessment of the DSM-5 level of personality functioning scale: The LPFS brief form (LPFS-BF). *Personality Disorders*, *7*(2), 192–197. doi:10.1037/per0000159

Hutsebaut, J., Kamphuis, J. H., Feenstra, D. J., Weekers, L. C., & De Saeger, H. (2016). Assessing DSM–5-Oriented level of personality functioning: Development and psychometric evaluation of the semi-structured interview for personality functioning DSM–5 (STiP-5.1). *Personality Disorders*, *8*(1), 94–101. doi:10.1037/per0000197

Jorm, A. F., Duncan-Jones, P., & Scott, R. (1989). An analysis of the re-test artefact in longitudinal studies of psychiatric symptoms and

personality. *Psychological Medicine*, *19*(2), 487–493. doi:10.1017/S0033291700012514

Karterud, S., Pedersen, G., Bjordal, E., Brabrand, J., Friis, S., Haaseth, O., … Urnes, O. (2003). Day treatment of patients with personality disorders: Experiences from a Norwegian treatment research network. *Journal of Personality Disorders*, *17*(3), 243–262. doi:10.1521/pedi.17.3.243.22151

Karterud, S., & Wilberg, T. (2007). From general day hospital treatment to specialized treatment programmes. *International Review of Psychiatry*, *19*(1), 39–49. doi:10.1080/09540260601080821

Kraemer, H. C., Kupfer, D. J., Narrow, W. E., Clarke, D. E., & Regier, D. A. (2010). Moving toward DSM-5: The field trials. *American Journal of Psychiatry*, *167*(10), 1158–1160. doi:10.1176/appi.ajp.2010.10070962

Krueger, R. F., Hopwood, C. J., Wright, A. G., & Markon, K. E. (2014). Challenges and strategies in helping the DSM become more dimensional and empirically based. *Current Psychiatry Reports*, *16*(12), 515. doi:10.1007/s11920-014-0515-3

Langbehn, D. R., Pfohl, B. M., Reynolds, S., Clark, L. A., Battaglia, M., Bellodi, L., … Links, P. (1999). The iowa personality disorder screen: Development and preliminary validation of a brief screening interview. *Journal of Personality Disorders*, *13*(1), 75–89. doi:10.1521/pedi.1999.13.1.75

Livesley, J. (2012). Tradition versus empiricism in the current DSM-5 proposal for revising the classification of personality disorders. *Criminal Behaviour and Mental Health*, *22*(2), 81–90. doi:10.1002/cbm.1826

Mancke, F., Herpertz, S. C., & Bertsch, K. (2015). Aggression in borderline personality disorder: A multidimensional model. *Personality Disorders*, *6*(3), 278–291. doi:10.1037/per0000098

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30. doi:10.1037/1082-989X.1.1.30

Morey, L. C. (2017). Development and initial evaluation of a self-report form of the DSM-5 level of personality functioning scale. *Psychological Assessment*, *29*(10), 1302–1308. doi:10.1037/pas0000450

Morey, L. C., Bender, D. S., & Skodol, A. E. (2013). Validating the proposed diagnostic and statistical manual of mental disorders, 5th edition, severity indicator for personality disorder. *Journal of Nervous and Mental Disease*, *29*(10), 729–735. doi:10.1097/NMD.0b013e3182a20ea8

Morey, L. C., Benson, K. T., Busch, A. J., & Skodol, A. E. (2015). Personality disorders in DSM-5: Emerging research on the alternative model. *Current Psychiatry Reports*, *17*(4), 558. doi:10.1007/s11920-015-0558-0

Narrow, W. E., Clarke, D. E., Kuramoto, S. J., Kraemer, H. C., Kupfer, D. J., Greiner, L., & Regier, D. A. (2013). DSM-5 field trials in the United States and Canada, Part III: Development and reliability testing of a cross-cutting symptom assessment for DSM-5. *American Journal of Psychiatry*, *170*(1), 71–82. doi:10.1176/appi.ajp.2012.12071000

OPD Task Force (Ed.) (2008). *Operationalized Psychodynamic Diagnosis OPD-2: Manual of diagnosis and treatment planning*. Ashland, OH: Hogrefe Publishing.

Oyserman, D. (2001). Self-concept and identity. In A. Tesser & N. Schwarz (Eds.), *Blackwell Handbook of Social Psychology: Intraindividual Processes* (pp. 499–517). Malden, MA: Blackwell. doi:10.1002/9780470998519.ch23

Pilkonis, P. A., Hallquist, M. N., Morse, J. Q., & Stepp, S. D. (2011). Striking the (Im)Proper balance between scientific advances and clinical utility: Commentary on the DSM-5 proposal for personality disorders. *Personality Disorders*, *2*(1), 68–82. doi:10.1037/a0022226

Pincus, A. L. (2011). Some comments on nomology, diagnostic process, and narcissistic personality disorder in the DSM-5 proposal for personality and personality disorders. *Personality Disorders*, *2*(1), 41–53. doi:10.1037/a0021191

Shedler, J., Beck, A., Fonagy, P., Gabbard, G. O., Gunderson, J., Kernberg, O., … Westen, D. (2010). Personality disorders in DSM-5. *American Journal of Psychiatry*, *167*(9), 1026–1028. doi:10.1176/appi.ajp.2010.10050746

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (1998). The mini-international neuropsychiatric interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59* (Suppl 20), 22–33;quiz 34–57.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi:10.1037/0033-2909.86.2.420

Silberschmidt, A., Lee, S., Zanarini, M., & Schulz, S. C. (2015). Gender differences in borderline personality disorder: Results from a multinational, clinical trial sample. *Journal of Personality Disorders*, *29*(6), 828–838. doi:10.1521/pedi_2014_28_175

Skodol, A. E. (2012). Personality disorders in DSM-5. *Annual Review of Clinical Psychology*, *8*, 317–344. doi:10.1146/annurev-clinpsy-032511-143131

Skodol, A. E., Morey, L. C., Bender, D. S., & Oldham, J. M. (2013). The ironic fate of the personality disorders in DSM-5. *Personality Disorders*, *4*(4), 342–349. doi:10.1037/per0000029

Skodol, A. E., Morey, L. C., Bender, D. S., & Oldham, J. M. (2015). The alternative DSM-5 model for personality disorders: A clinical application. *American Journal of Psychiatry*, *172*(7), 606–613. doi:10.1176/appi.ajp.2015.14101220

Thylstrup, B., Simonsen, S., Nemery, C., Simonsen, E., Noll, J. F., Myatt, M. W., & Hesse, M. (2016). Assessment of personality-related levels of functioning: A pilot study of clinical assessment of the DSM-5 level of personality functioning based on a semi-structured interview. *BMC Psychiatry*, *16*, 298. doi:10.1186/s12888-016-1011-6

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The prevalence of personality disorders in a community sample. *Archives of General Psychiatry*, *58*(6), 590–596. doi:10.1001/archpsyc.58.6.590.

Trevethan, R. (2017). Intraclass correlation coefficients: Clearing the air, extending some cautions, and making some requests. *Health Services and Outcomes Research Methodology*, *17*(2), 127–143. doi:10.1007/s10742-016-0156-6

Tyrer, P., Crawford, M., & Mulder, R. (2011). Reclassifying personality disorders. *Lancet*, *377*(9780), 1814–1815. doi:10.1016/s0140-6736(10)61926-5

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *Lancet*, *385*(9969), 717–726. doi:10.1016/s0140-6736(14)61995-4

Urbaniak, G. C., & Plous, S. (1997). Research randomizer. Retrieved July, 7, 2008.

Verheul, R., Andrea, H., Berghout, C. C., Dolan, C., Busschbach, J. J., van der Kroft, P. J., … Fonagy, P. (2008). Severity indices of personality problems (SIPP-118): Development, factor structure, reliability, and validity. *Psychological Assessment*, *20*(1), 23–34. doi:10.1037/1040-3590.20.1.23

Verheul, R., Bartak, A., & Widiger, T. (2007). Prevalence and construct validity of personality disorder not otherwise specified (PDNOS). *Journal of Personality Disorders*, *21*(4), 359–370. doi:10.1521/pedi.2007.21.4.359

Waugh, M. H., Hopwood, C. J., Krueger, R. F., Morey, L. C., Pincus, A. L., & Wright, A. G. C. (2017). Psychological Assessment with the DSM-5 Alternative Model for Personality Disorders: Tradition and Innovation. *Professional Psychology, Research and Practice*, *48*(2), 79–89. doi:10.1037/pro0000071

Weertman, A., Arntz, A., Dreessen, L., van Velzen, C., & Vertommen, S. (2003). Short-interval test-retest interrater reliability of the Dutch version of the Structured Clinical Interview for DSM-IV personality disorders (SCID-II). *Journal of Personality Disorders*, *17*(6), 562–567. doi:10.1521/pedi.17.6.562.25359

Westen, D., & Arkowitz-Westen, L. (1998). Limitations of axis II in diagnosing personality pathology in clinical practice. *American Journal of Psychiatry*, *155*(12), 1767–1771. doi:10.1176/ajp.155.12.1767

Widiger TA, Livesley W.J (2001). Official classification systems. *Handbook of personality disorders*. (pp. 60–83). New York: Guilford.

Wilkinson-Ryan, T., & Westen, D. (2000). Identity disturbance in borderline personality disorder: An empirical investigation. *American Journal of Psychiatry*, *157*(4), 528–541. doi:10.1176/appi.ajp.157.4.528

Zachar, P., Krueger, R. F., & Kendler, K. S. (2016). Personality disorder in DSM-5: An oral history. *Psychological Medicine*, *46*(1), 1–10. doi:10.1017/s0033291715001543

Zimmerman, M. (2013). What is ironic about wanting empirical support to justify changes in diagnostic criteria? Commentary on "the ironic fate of the personality disorders in DSM-5". *Personality Disorders*, *4*(4), 352–353. doi:10.1037/per0000048

Zimmerman, M., Chelminski, I., Young, D., Dalrymple, K., & Martinez, J. (2012). Impact of deleting 5 DSM-IV personality disorders on prevalence, comorbidity, and the association between personality disorder pathology and psychosocial morbidity. *The Journal of Clinical Psychiatry*, *73*(2), 202–207. doi:10.4088/JCP.11m07140

Zimmermann, J., Benecke, C., Bender, D. S., Skodol, A. E., Schauenburg, H., Cierpka, M., & Leising, D. (2014). Assessing DSM-5 level of personality functioning from videotaped clinical interviews: A pilot study with untrained and clinically inexperienced students. *Journal of Personality Assessment*, *96*(4), 397–409. doi:10.1080/00223891.2013.852563

Zimmermann, J., Ehrenthal, J. C., Cierpka, M., Schauenburg, H., Doering, S., & Benecke, C. (2012). Assessing the level of structural integration using operationalized psychodynamic diagnosis (OPD): Implications for DSM-5. *Journal of Personality Assessment*, *94*(5), 522–532. doi:10.1080/00223891.2012.700664